Predicting Major League Baseball Strikeout Rates from Differences in Velocity and Movement Among Player Pitch Types

By

Eric P. Martin

Thesis Project
Submitted in partial fulfillment of the
Requirements for the degree of

MASTER OF SCIENCE IN PREDICTIVE ANALYTICS

December, 2018

Thomas W. Miller, First Reader

Bradley R. Smith, Second Reader

ABSTRACT

Predicting Major League Baseball Strikeout Rates from Differences in Velocity and Movement Among Player Pitch Types

Eric P. Martin

Despite several studies examining how fastball velocity, curveball movement, or other pitch characteristics affect pitching results, no analysis has examined pitching performance based on velocity and movement relationships among all pitch types in a pitcher's arsenal. Using PITCHf/x information for over 2.5 million Major League Baseball pitches thrown by 402 different pitchers from 2012 through 2017 this thesis attempts to answer the question: "Can a model effectively predict a pitcher's strikeout rate by examining differences in velocity and movement among pitch types?"

Each player's pitch types are identified using hierarchical clustering and an adjusted Bayesian Information Criterion calculation. Several supervised and unsupervised models predict pitcher strikeout rates by examining differences in velocity, movement, release points, pitch frequency, and the Mahalanobis distance between pitch types. A random forest model has the lowest error rate with a 2.93% mean absolute error of prediction (MAE) and 86.5% of the predictions within 5.0% of the actual strikeout rates. The MAE is 0.37 points lower (i.e., 2.94% vs. 3.31%) than simply using fastball velocity and strike rate to predict strikeout percentages.

Pitch velocity and vertical movement have the most significant impact on a pitcher's strikeout percentages. Understanding these performance drivers enables players and coaches to target the elements most likely to increase strikeouts and identify promising young pitchers with otherwise unremarkable pitching statistics.

Acknowledgements

When I began this analysis in January 2018, I was blissfully unaware of the challenges I would encounter. The basic task of identifying pitch types took considerable effort to complete and identifying a method to quantify differences between pitch types was challenging. Through brute force I was able to push through these challenges after taking an innumerable number of fruitless paths. It was a valuable endeavor, however, as the process led to a much better understanding of statistical models and the R programming language.

I would like to thank Prof. Tom Miller for his recommendations on simplifying the analysis, measuring pitch entropy, and developing a broad variance measure applicable to pitchers having differing numbers of pitch types. I would also like to thank Prof. Brad Smith for his insights and suggestions for future analyses. Finally, this thesis would not have been possible without the support and patience of my wife. She allowed me to pursue this project at my own pace, desiring that I receive the most out of this exercise. Her unwavering willingness to also serve as a sounding board for my brainstorming sessions were more valuable than she realizes. I am forever grateful.

Table of Contents

ABSTRACT 3

Acknowledgements 4

Introduction 9

Statement of the Problem 9

Justification 9

Literature Review 10

Performance Measurement 10

Pitch Factors Affecting Strikeouts 12

Measuring Deception 14

Methods 16

Data Sourcing 16

Statistics By Season 17

Qualifying Pitchers 17

Metrics Examined 19

Identifying Pitch Types – Clustering 23

Distance Between Pitch Clusters 28

Independent Variables 33

Models Examined 37

Results 39

Pitch Types 39

Model Results 40

Effect of MLB Strikeout Rate Increases on Prediction Errors 43

Conclusions 45

List of Figures

Figure 1. Euclidean distance between pitch pairs by a right-handed pitcher. 21

Figure 2. Euclidean distance between pitch pairs by a left-handed pitcher. 21

Figure 3. Max Scherzer 2016 pitch clusters as identified by PITCHf/x. 24

Figure 4. Max Scherzer 2016 pitch clusters as calculated in this analysis................................ 24

Figure 5. Percentage of pitches removed as outliers from observations in the data set. 27

Figure 6. Mean pitch velocities and vertical movements for right-handed pitchers in 2016. 29

Figure 7. Pitch velocity and horizontal movement by left-handed pitchers from 2012−16......... 31

Figure 8. Pitch velocity and horizontal movement by right-handed pitchers from 2012−16. 31

Figure 9. WMD-1 against strikeout percentage from 2012−16... 34

Figure 10. WMP-BP against strikeout percentage from 2012−16. 34

Figure 11. Strikeout percentage by number of pitch types from 2012−16. 39

Figure 12. Random forest model variable importance plot (top ten variables). 41

Figure 13. Random forest model 2017 predicted strikeout percentages against 2017 actual values. 42

Figure 14. Percentage of random forest model predictions within error ranges. 42

Figure 15. MLB annual strikeout percentages from 2005–2017... 43

Figure 16. Distribution of strikeout percentages in the training (2012−16) and test (2017) sets. 43

List of Tables

Table 1. Number of Pitchers with 3–9 Pitch Clusters in the Training and Test Sets 28

Table 2. Details of the Three Colored points in Figure 6..29

Table 3. Velocity and Movement Correlation Matrixes for Left and Right-Handed Pitchers.....32

Table 4. Statistics for the 894 Training Set Observations Used to Create Models Observations 36

Table 5. Test Set Errors for All Models...40

Table 6. Top 10 Random Forest Model Predictions..42

Table 7. Top Strikeout Percentage Predictor Variables...44

List of Equations

Equation 1. Converting 2017 PITCHf/x vector velocities into pre-2017 version of initial velocity ………………………………………………………………………..……………. 20

Equation 2. Adjusted Bayesian Information Criterion used to identify optimal number of pitch clusters ………………………………………………………………………………… 26

Equation 3. Mahalanobis distance calculation ……………………………………………… 31

Equation 4. Weighted Mahalanobis Distance From 1 ……………………………………… 33

Equation 5. Weighted Mahalanobis Distance Between Pairs ………………………………. 33

Equation 6. Entropy ………………………………………………………………………… 34

Introduction

Statement of the Problem

Many factors affect pitching success in Major League Baseball (MLB). Pitch velocity, movement, and accuracy are a few of the most important pitch characteristics impacting a pitcher's ability to induce outs. Baseball enthusiasts have frequently examined how individual pitch characteristics like fastball velocity and curveball movement affect pitching results. Studying single pitch characteristics, however, fails to account for relationships between different pitches thrown by a pitcher. Few analyses have examined the interdependence of velocity and movement among pitches and their effects on pitching performance.

The relationship between a pitcher's various pitches is commonly referred to as a pitcher's "stuff." The ability to effectively change speeds and movement is central to having good "stuff." Branch (2015) observed that although a pitcher's "stuff" is one of the most oft-used terms to describe pitching ability, it has no established definition. According to Branch (2015), "Stuff can describe a collection of pitches, how well those pitches are thrown on a particular day, and how well they fool hitters" (p. A1). This interpretation is decidedly subjective and eludes consensus. Scouting effective pitchers based on an amorphous recognition of "stuff" quickly devolves into an abstract exercise. This thesis attempts to isolate the primary elements of a pitcher's "stuff" and model their impact on performance.

Justification

Baseball statisticians have not yet identified a way to predict pitching success by examining the disparate speeds and movements of all of the pitches in a pitcher's arsenal. This analysis proposes a method for doing so. This thesis predicts pitcher strikeout rates by exploring

(1) horizontal and vertical pitch movements, (2) the difference between pitch speeds, and (3) the number of distinct pitches thrown by pitchers. This analysis attempts to answer the question: "Can a model effectively incorporate pitch speed, movement, and variety to predict MLB pitcher success?"

Statisticians have recognized that fastball velocity and strike rates affect pitchers' strikeout percentages. (See Cameron, 2009; Arthur, 2014). A linear regression model using these two variables to predict strikeout percentage has an adjusted r-squared value of 0.230 and, on average, strikeout percentage predictions are within 3.31 percentage points of their actual values (MAE = 0.0331). This analysis improves on existing analyses by including differences in pitch velocity and movement among pitches to improve the accuracy of these predictions. Understanding these impacts may help players and coaches improve success by making changes to pitching approaches. It may also enable scouts to identify promising young pitchers who otherwise have unremarkable pitching statistics.

Literature Review

Performance Measurement

Many metrics measure pitcher performance, the most familiar of which is a pitcher's earned run average (ERA). Among the many pitching statistics, however, there is no single statistic that "best" measures performance success. Albert (2006) examined several pitching rate statistics to determine the most reliable measure of performance. Albert used a random effects model to isolate the causes of statistical variation. The study revealed a pitcher's strikeout rate—the number of strikeouts divided by the number of batters faced—is the statistic least likely to be affected by chance. Not only is strikeout percentage agnostic to a team's defensive ability, it can be remarkably consistent from year to year (Albert, 2006). Woolner (2006) stated a pitcher's strikeout percentage is one of “the most reliable and consistent pitching statistics” (p. 56). Moreover, given its low year-over-year variance, differences in a pitcher’s annual strikeout percentages can more confidently be attributed to changes in pitch characteristics (e.g., movement, velocity, and accuracy) rather than other factors. Strikeout percentage is used in this analysis to measure pitching performance due to its simplicity, reliability, and independence from fielding events.

Strikeout percentage is chosen over other common pitching performance statistics. For example, fielding independent percentage (FIP) builds on a pitcher’s strikeout rate by adding a weighted mix of the rate of home runs, walks, and hit by pitches per inning.[1] (Albert, 2016). FIP attempts to account for results within the pitcher’s exclusive control by ignoring runs scored and outs. It nevertheless suffers from several factors that increase variance. For example, like ERA, FIP measures performance on a per inning, rather than per batter, basis. Since almost any number of batters can appear in an inning, pitchers with similar FIP statistics can have dramatically different performances.

Wins above replacement (“WAR”) is another commonly used statistic to measure pitching performance. It has been described as “a way of adding up the various measurable components of a player’s performance, setting a standard baseline, and using it to compare player value across seasons, leagues, even across eras” (Law, 2017, p. 202). A significant impediment to using this statistic to measure performance is that statisticians calculate WAR in various ways. Many calculation methodologies are proprietary and not publicly available. For

[1] FIP = (13HR + 3(BB+HBP) - 2SO)/IP + Constant. (Fangraphs, 2015b). The constant is an annual adjustment to enhance interpretation by ensuring the league-average FIP and ERA statistics are approximately equal. (Albert, 2016).

these reasons, some commentators have objected to using WAR to measure performance given WAR's various definitions and lack of reproducibility (Baumer, Jensen, & Matthews, 2015).

Pitch Factors Affecting Strikeouts

Several pitch attributes are strongly correlated with strikeout percentage. Most fans recognize the correlation between high strikeout rates and pitchers with fastballs near or exceeding 100 mph. Indeed, Cameron (2009) and Arthur (2014) found a positive correlation between a pitcher's maximum pitch velocity (i.e., fastball velocity) and his strikeout rate. This makes sense given the short reaction time hitters have to determine whether, when, and where to swing (Gray, 2002).

Rarely will exceptional velocity alone be enough to generate strikeouts. Pitchers deceive hitters by changing the speed, movement, location, and sequence of pitches. More than the absolute velocity or movement of a particular pitch, the difference in velocity and movement between pitches is a primary basis for deception. A curveball is more effective if its movement is dramatically different from a fastball's movement. These pitches are interdependent on each other. It is relationships among all pitches in a pitcher's arsenal—rather than simply the effectiveness of a single pitch—that determines success (Sarris, 2018).

In some instances, a potent combination of pitches can be more effective than a high velocity fastball. Hall of Fame pitcher, Greg Maddux, is a prime example. Maddux won four Cy Young awards despite having a fastball in the lows 90s (Davis, 2016). The movement of his pitches was able to consistently keep batters off-balance (Davis, 2016). More recently, Rescan (2017) summarized this phenomenon in his article descriptively titled: "Kyle Hendricks's greatness is about more than control and command: His velocity-less success depends on the movement, too." Rescan (2017) found the difference in movement among Hendricks' pitches to be a primary element in deceiving batters, despite Hendricks's relatively low pitch velocity.

The present analysis looks beyond single pitch characteristics and seeks to understand how relationships among pitches induce strikeouts. Work on this subject has been more limited. Professor Rob Gray conducted one of the first experiments in this area in 2002. He developed a virtual batting simulation to test temporal and spatial swing accuracy against a variety of pitches. Using college baseball players as test subjects, Gray (2002) found hitting was "nearly impossible in a situation where pitch speed is random and in which no auxiliary cues (e.g., pitcher's arm motion or pitch count) are available to the batter" (p.1140). Gray compared these results to a control scenario where batter swing accuracy was significantly better when seeing only two pitch speeds. Batters suffered (and pitchers benefited) as the number of different pitch speeds increased. These results led Gray to conclude, "it is important for baseball pitchers to learn to throw at least three different types of pitches" (Gray, 2002, p. 1143). See also Ryan and House (1991).

Gray's work is somewhat limited, however, in that it tested only six different college batters, each receiving only 20 pitches per experiment. This was a small and unrepresentative sample, especially when compared to the experience level and number of at bats taken by MLB batters. Gray's analysis was nevertheless instrumental in its use of an actual batting simulation to isolate specific temporal and spatial pitch effects and provided a foundation for future work.

Subsequent analyses using MLB PITCHf/x data examined swing-and-miss rates over various velocities and pitch movements. Hale (2013) discovered swings and misses occur most often when pitches have high velocity and low vertical movement. Surprisingly, swing-and-miss rates generally ***decreased*** as vertical movement increased—even at high velocities.

Measuring Deception

Gray (2002) demonstrated that batters have difficulty contacting pitch combinations having several different speeds. Hale (2013) extended Gray's findings to MLB players and included the interaction between pitch velocity and movement when examining swinging strike rates. Other analyses examined how changes in movement or velocity between consecutive pitches affect swinging strike or strikeout rates. Specifically, Roegele (2014) examined swinging strike rates based on the distance between consecutive pitches measured at both home plate and 33 feet from home plate. Swinging strikes were related not only to the distance between pitches when crossing home plate, but also their separation at 33 feet. Pitch pairs that looked the same at 33 feet cause swinging strikes when separated by only a few inches when crossing home plate. Carleton (2015) calculated how velocity differences between pitch pairs affected batter contact rates. Batters had a spectrum of different contact rates for pairs of pitches with different speeds, although Carleton did not provide any universal conclusions connecting velocity differences and contact rates. Bonney (2015) discovered that a pitch thrown more than 5 mph slower than the previous pitch reduced expected runs. This effect was greater when the first of the two consecutive pitches was a ball rather than a strike. None of these studies examined a pitcher's entire arsenal of pitches, however, or how several velocities or movements between pitches have a compounding effect on each other.

Other attempts have been made to quantify how a pitcher's pitch collection induces outs. Sarris (2014), Schwartz (2014), and Jackman (2015) developed various "arsenal scores" to evaluate the effectiveness of a pitch collection. Scores were calculated using the swinging strike rate, groundball rate, zone percentage rate, and/or pitch usage rate for each pitch type in a pitcher's collection. While the methods varied across researchers, each calculation essentially summed the impacts of individual pitch types to create an aggregate score. Unlike the present analysis, these studies did not specifically examine the interactions between pitch types or explore how pitch characteristics such as velocity or movement affect strikeout rates.

Pitch deception is caused by the uncertainty in a pitch's velocity and movement. Shannon (1948) introduced an entropy metric to quantify uncertainty by examining probabilities of potential outcomes. Arthur (2014) used these entropy principles to quantify pitch uncertainty based on the number of pitch-types and the "evenness," or uniformity, of pitch distribution. In doing so, Arthur found that increased entropy is correlated with larger strikeout rates. The present analysis includes an entropy metric to quantify uncertainty resulting from the number of pitch types and the frequency with which each is thrown.

More recent work has examined the concept of pitch tunnels and their impact on pitcher strikeout rates. (See Pavlidis, Judge, & Long, 2017). Pitch tunnels measure the difference in pitch positions at the hitter's decision-making point (about 23.8 feet from home plate) while considering the pitcher release point, pitch sequencing, break location, and velocity (Pavlidis et al., 2017). This thesis builds on aspects of this pitch tunnel research by examining how differences in pitch release points affect strikeout rates.

Most germane to this analysis, Healey, Zhao, and Brooks (2017c) developed a pitch sequencing model examining pitcher strikeout rates as a function of pitch velocity and movement. For every pair of consecutive pitches, they calculated correlation coefficients for horizontal and vertical pitch location, horizontal and vertical pitch movement, and pitch velocity. High correlation coefficients indicated a pitcher's tendency to throw similar pitches consecutively. Their study used these correlation coefficients to model pitcher strikeout rates.

Somewhat contrary to findings by Hale (2013), Healey et al. (2017c) discovered pitchers' strikeout rates increased when fastball velocity and vertical movement increased. Healey et al.

(2017c) also showed strikeout rates decrease when fewer fastballs are thrown and when pitch-to-pitch correlations increase. The study's conclusion suggested "a more detailed model could include information about the number, frequency, and physical properties of a pitcher's off-speed pitches and how well these pitches complement each other and the pitcher's fastball" (p. 101). That proposition was the catalyst for this thesis.

The present analysis expands upon previous studies in several ways. Most significantly, many of the prior analyses only examined correlations between pairs of consecutive pitches. Researchers did not provide an omnibus measure for all pitches thrown by a pitcher like the one set forth in the present analysis. This thesis attempts to quantify a pitcher's "stuff" by understanding holistic relationships across all pitch types.

Methods

This analysis collects information on over 4.3 million pitches thrown by 2,093 MLB pitchers from 2012 through 2017. As described below, model-based clustering techniques use pitch velocities, horizontal movements, and vertical movements to group similar pitch types. A Mahalanobis distance measurement calculates the difference between pitch types by accounting for correlation between these variables. Several statistical models, employing supervised and unsupervised machine learning techniques, are created to predict pitcher strikeout rates. Models are tested and compared based on MAE rates.

Data Sourcing

The pitchR/x package in R is used to collect PITCHf/x data for the 2012–2017 baseball seasons from MLB Advanced Media's Gameday database. (Sievert, 2014). This database includes extensive information on every pitch: release point, vertical and horizontal movement, velocity, location when crossing the plate, etc. (Kagan, 2009; Fast, 2010a; Sievert, 2014). Data is cleaned by deleting 13,815 observations/pitches with missing velocity, horizontal movement, or vertical movement values. An additional 46,940 observations with null or unidentified pitches are also deleted. 4,252,170 pitches remain in the final dataset. MLB pitching statistics from each season are loaded into R as CSV files from Baseball-Reference.com (Sports Reference LLC, 2018) and merged with the consolidated pitch information for each qualifying pitcher.

The data is separated into training and test sets to analyze model accuracy. Models are built using the training set which includes data from the 2012–2016 seasons. Data from the 2017 season is used as the test set to evaluate the accuracy of the final models.

Statistics By Season

Rather than using batter-level or game-level pitching results, this analysis examines pitch information from full MLB seasons (typically between 1,500 and 3,500 pitches each for starting pitchers). Using a full season attempts to normalize the talent of opposing batters since pitchers will generally face a more even mix of talent over the course of an entire season. This allows for more accurate comparisons across pitchers. Similarly, using a full season smooths differences in pitch strategies against left and right-handed batters thereby eliminating this potentially confounding variable. Finally, there are well-documented measurement discrepancies across stadiums due to minor systematic errors in each stadium's PITCHf/x and Statcast systems (Fast, 2010b; Fast, 2011; Garik, 2011; Marchi, 2011; Boyle, O'Rourke, Long, & Pavlidis, 2018; Arthur, 2017; Kagan, 2018). Examining a full season of pitches reduces the impact of these measurement discrepancies.

Qualifying Pitchers

To control for biased pitching results, short relief pitchers are excluded from the analysis. Measures of pitching effectiveness of many relievers are distorted because they often enter games to face batters in favorable match-ups. In these cases, pitchers pitch to a small number of batters where the pitchers' odds of success are maximized—for example, left- or right-handed "platoon" situations. These circumstances positively bias a pitcher's statistics (Albert, 2006). Pitchers who face a small number of batters per appearance are therefore excluded from the analysis to provide a more reliable representation of each pitcher's abilities.

Separating short-relief pitchers from starting and long-relief pitchers requires certain assumptions. Although most teams have starting pitching rotations, these pitchers change frequently over the course of a season due to injury, ineffectiveness, and rest considerations. To qualify for an individual pitching champion award, MLB requires pitchers to pitch at least as many innings as the number of games played by his team. (MLB rule 9.22(b), Office of the Commissioner of Major League Baseball, 2016). This requires pitching 162 innings over a typical 162-game season; a fairly restrictive cutoff. This includes only 467 pitchers in the 6 seasons from 2012–2017—an average of fewer than 78 pitchers per season (Sports Reference LLC, 2018). Some analyses have lowered this threshold and exclude pitchers with fewer than 100 innings per season (Piette, Braunstein, McShane, & Jensen, 2010). Using the number of innings pitched as the sole qualifier, however, potentially includes heavily-used relievers regularly brought into games to face favorable match-ups. It is important to consider the average number of batters faced per pitching appearance to ensure a pitcher faces a representative number of batters in long-relief, spot starts, or seasons shortened by injury.

Two criteria are therefore used in this analysis to identify "qualifying pitchers": (1) the average number of batters faced per appearance and (2) the number of pitches thrown each season. In this analysis, pitchers who both faced an average of 10 batters per appearance and pitched at least 1,000 pitches in a season are "qualified." These qualifiers ensure inclusion of sample sizes suitable to identify each pitcher's pitch collection while reducing bias resulting from favorable match-ups.

Using the above criteria to identify qualifying pitchers, the training set contains 894 observations for 360 different pitchers from 2012–2016. 228 of these players pitched in two or more seasons. The 2017 test set includes 170 pitchers. Despite pitching in multiple years, several pitchers in the training set had year-over-year changes in pitch characteristics resulting in very different pitch profiles. These performance differences for the same pitcher are ideal for isolating how pitch movement and velocity differences affect strikeout percentages.

Metrics Examined

Although the PITCHf/x database includes over 70 variables for each pitch, this analysis uses only three measurements to identify pitch types: velocity, horizontal ball movement, and vertical ball movement. These metrics measure the essential pitch dynamics batters face when attempting to contact the baseball. As described in more detail below, pitch velocity is measured in miles per hour and ball movements are measured in inches (Fast, 2007).

Through 2016, PITCHf/x reported velocity at a point 50 feet from home plate (Cameron, 2017). Although a pitcher's typical release point is about 55 feet from home plate, a 50 foot measurement allowed for differences among pitcher release points. The velocity at 50 feet is less than the velocity at the release point, however, due to the baseball's natural deceleration. MLB changed the system used to measure PITCHf/x velocity in 2017 and began measuring velocity at a pitcher's exact release point (Cameron, 2017). The result was an increase in reported velocities (Cameron, 2017). For consistency in this analysis, 2017 reported velocities are adjusted to reflect velocities 50 feet from home plate. Using the method outlined by Nathan & Brooks (2017), initial velocities in 2017 are calculated as the square root of the sum of the initial horizontal (x), vertical (z), and incoming (y) vector velocities, with the result converted from feet per second to miles per hour. Since PITCHf/x still measured vector velocities at 50 feet from home plate in 2017, converting these vector velocities into a pre-2017 version of initial velocity (*v₀*) is straightforward (Nathan & Brooks, 2017):

$$For\ 2017{:}v_0 = \sqrt{v_x^2 + v_y^2 + v_z^2}\ x\ \left(\frac{3600}{5280}\right)\frac{mph}{ft/_{sec}} \qquad (1)$$

PITCHf/x measures a pitch's movement relative to a theoretical pitch thrown at the same speed with no spin-induced change in direction (Fast, 2007). In its simplest terms, movement is the "deviation of the trajectory from a straight line with the effect of gravity removed" (Nathan, 2012, p. 2). Although there is disagreement about how PITCHf/x calculates it (Nathan, 2012), movement captures the change in the baseball's location relative to a non-spinning object moving at the same velocity.

Vertical movement measures how much a pitch drops relative to a pitch with the same velocity without spin (gravity still causes the ball to move downward, but PITCHf/x ignores this effect). Zero vertical movement therefore means a pitch drops the same amount as a pitch without spin. (See Richmond, 2015). Pitches with negative vertical movement drop more than expected (e.g., curveballs) while those with positive values have significant backspin and fall less than expected—sometimes appearing to hang in mid-air. A pitch with positive vertical movement still moves downward, but does so less than a same-velocity ball without spin.

Horizontal movement measures the baseball's left and right change in direction from the catcher's perspective. Pitches with positive values move to the right of a theoretical "non-moving" pitch while those with negative values move to the left. To a right-handed batter, positive horizontal values “break away” and negative values “break in.” Positive vertical values “break up” while negative values “break down”.[2]

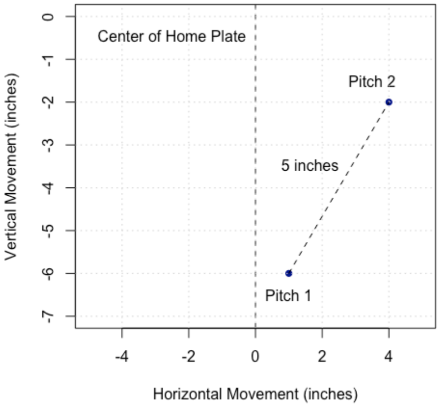

Figure 1. Euclidean distance between pitch pairs by a right-handed pitcher.

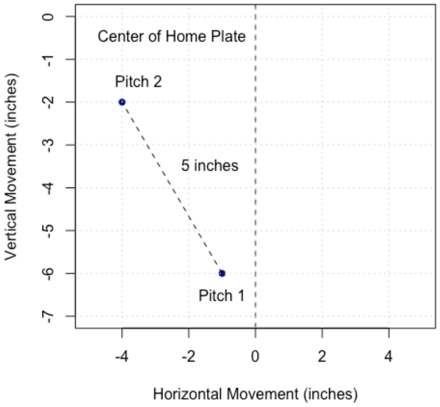

Figure 2. Euclidean distance between pitch pairs by a left-handed pitcher.

This analysis examines the relative relationship between pitch movements rather than the ordinal direction of movement. Measurements are therefore not separated based on the handedness of either the batter or the pitcher. For example, imagine a right-handed pitcher with two pitches: the first moving 1 inch right and 6 inches down and the second moving 4 inches right and 2 inches down (Figure 1). The Euclidean distance between these pitches is $\sqrt{(1-4)^2+(6-2)^2} = 5$ inches. Conversely, the natural horizontal movement of pitches from a left-handed pitcher is in the opposite direction. Comparable pitches from a left-handed pitcher would move in the opposite horizontal direction but would also be 5 inches apart (Figure 2). As this analysis focuses on the distance between pitches (5 inches), rather than their absolute direction, the handedness of the pitcher is not considered. Pitcher handedness is only considered when calculating Mahalanobis distances between pitches as discussed in detail below.

[2] A helpful video visualizing differences in pitch movement is at http://media.giphy.com/media/ytGHRSRknbujm/giphy.gif.

In addition to metrics measuring pitch movement and velocity, pitcher release points are included when modeling to predict strikeout percentages. Differences in release points can serve to signal pitch types to batters and diminish the element of deception. (Gray, 2002). Right-handed and left-handed pitchers typically have negative and positive x-values, respectively, since the horizontal release point is measured relative to the middle of home plate. The vertical release point is the distance from the ground. Both metrics are measured in feet. Differences between release points used in this analysis are always positive values.

Two additional variables are included in the models to minimize the impact of factors unrelated to changes in the velocity and movement between pitches. First, pitcher strike rates are included as a way to normalize accuracy differences between pitchers. Pitchers with the same pitch characteristics may have different strikeout rates simply because one pitcher throws more strikes than the other. Second, models consider whether pitchers played in the National League, American League, or both each season. National League pitchers may have larger strikeout rates because they face opposing pitchers who take one or more at bats each game.

Other pitch measurements were considered, but not used in this analysis for several reasons. For example, reports have examined the rate and direction of spin on the baseball to predict pitching performance. (See Perpetua, 2016). A baseball's spin creates a Magnus force, which in turn results in the ball's movement. (Adair, 1995; Nathan, 2012) Spin rate is therefore a proxy for horizontal and vertical movement. Spin and movement are the cause and effect of the same phenomena. Despite this, the same spin does not always result in the same movement. Temperature, humidity, wind speed, and altitude have various impacts on ball movement. (See

Adair, 1995; Florio & Shapiro, 2016). By examining a baseball's actual vertical and horizontal movement, this analysis ignores these confounding variables on a baseball's spin rate.[3]

A pitch's break-point was another variable considered. Some commentators refer to pitches as having "late-break," referring to baseballs that move later in their flight path than traditional pitches (Nathan, 2013, para. 1). The concept of a late-break is a misnomer, however. Professor Alan Nathan at the University of Illinois explained that since a pitch's break is continuous, starting from when the ball is released, "late break actually means 'not much break,' meaning that the difference between no break and actual break is small enough at 20 feet [the point when a batter initiates his swing] that the batter can't perceive it" (Nathan, 2013, para. 6). For this reason, pitch break points are not included in the analysis.

Finally, this thesis focuses on the relationship between each pitcher's collection of pitches. It ignores broader pitching effects that induce outs. For example, pitch location and sequencing are not evaluated here, although they undoubtedly affect pitcher performance. (See Gray, 2002; Healey, Zhao, & Brooks, 2017a; Sidle & Tran, 2018; Trueblood, 2018).

Identifying Pitch Types – Clustering

Every pitcher throws several types of pitches. Identifying these pitches can be a difficult task. For example, a sinker from one pitcher may move the same as another pitcher's four-seam fastball. (See Healey et al., 2017a). Baseball enthusiasts have made considerable efforts to label pitch types and Statcast and PITCHf/x label every pitch using a proprietary neural net classification algorithm developed by MLBAM based on several metrics including velocity, spin rate, and movement (Fast, 2010a).

[3] Although some batters can anticipate pitch movements by observing an approaching ball's spin rate and direction, not all batters see the ball's spin. (See Schmidt & Ellis, 1994). Regardless, movement differs based on the environment and game conditions (Gray, 2002).

PITCHf/x pitch type designations are sometimes inconsistent and can label the same ball movement differently among pitchers (Fast, 2010b). These pitch designations are not reliable for use in a robust pitch analysis (Fast, 2010b). Relying on PITCHf/x pitch type designations to make predictions based on changes in movement and velocity can lead to unreliable results. Nevertheless, whether a pitch is labeled a sinker or a four-seam fastball is irrelevant to a hitter. Hitters are simply concerned with pitch speed and movement. What matters is a pitcher's ability to deceive a hitter by changing the velocity, movement, and location of each pitch.

This analysis therefore determines each pitcher's pitch repertoire by using a model-based clustering algorithm instead of PITCHf/x pitch classifications. Specifically, the multivariate Gaussian mixture model introduced by Pane, Ventura, Steorts, and Thomas (2013) is used in this analysis to cluster pitch types. Pane et al. (2013) determined that model-based clustering more accurately identifies MLB pitch types than either k-means or neural network clustering. Model-based clustering is better able to identify pitches with differing variances and often reduces the number of small clusters.

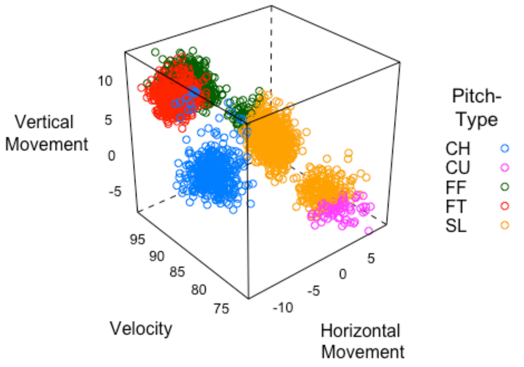

Figure 3. Max Scherzer 2016 pitch clusters as identified by PITCHf/x. Five pitch types are identified: change-up (CH), cutter (CU), four-seam fastball (FF), two-seam fastball (FT), and slider (SL).

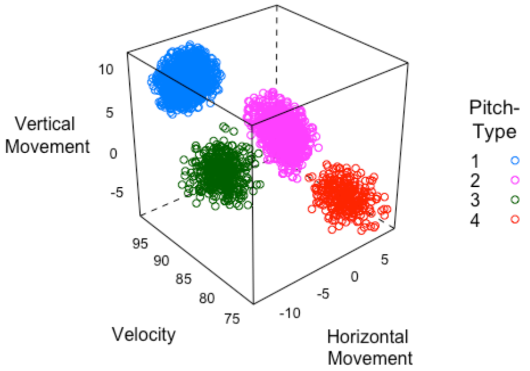

Figure 4. Max Scherzer 2016 pitch clusters as calculated in this analysis. Four clusters identified corresponding to four pitch types.

Figure 3 demonstrates potential problems with relying on PITCHf/x designations to identify pitch types. The PITCHf/x labels have significant overlap between comparable pitches. The clustering technique used in this analysis instead groups similar pitches together while removing outliers. Figure 4 shows the same 2016 Max Scherzer pitches clustered using the methods outlined in this analysis.

The mclust package in R is used to assign pitches to clusters (Fraley et al., 2012). To simplify the analysis, only velocity, horizontal movement, and vertical movement variables are included when clustering pitches. Pitches are assigned to clusters using an agglomerative hierarchical clustering method based on maximum likelihood criteria for parameterized Gaussian mixture models (Evans, Love, Thurston, 2015; Fraley & Raftery, 2002; Fraley et al., 2012). Evans et al. (2015) describe the algorithm's steps:

> From the initial clusters, Maximum Likelihood Estimation is carried out via the Expectation-Maximization (EM) algorithm for parameter estimation. Until convergence occurs, the EM algorithm iterates between the 'E'-step, which computes z_{ik}, the conditional probability that observation *i* belongs to group *k* given the current parameter estimates and the 'M'-step, which computes parameter estimates given *z*. Once estimation is completed for each specified number of clusters and cluster variance structure, a final model (defined by the number of clusters and the cluster variance structure) is selected, based on the largest BIC value. (Evans et al., 2015, p. 67).

Each pitcher's pitches are separated into nine alternative cluster configurations: from 1 to 9 pitch clusters. The number of clusters with the lowest adjusted Bayesian Information Criterion (BIC_{adj}) are identified for each pitcher. Using the methods employed by Pane et al.

(2013), BIC values for each number of clusters are adjusted using penalties based on the number of clusters and high intra-cluster correlation coefficients. This reduces the number of small clusters and provides a more reliable representation of each pitcher's pitch types.

$$\mathrm{BIC}_{\mathrm{adj}} = \underbrace{-2 \log (f(Y \mid \hat{p})}_{\text{BIC}} - \underbrace{2\lambda \textstyle\sum_i log\, (f(c_i))}_{\text{1st Penalty}} + \underbrace{[k \cdot (j + \frac{j(j\text{-}1)}{2} + (k-1))] \cdot \log(n)}_{\text{2nd Penalty}} \tag{2}$$

Equation 2 provides the adjusted Bayesian Information Criterion used to identify the optimal number of pitch clusters. It includes the likelihood function ($f(Y \mid \hat{p})$), the number of clusters (k), the sum of the upper off-diagonal elements of a correlation matrix ($\sum_i f(c_i)$), and the total number of pitches (n) (Pane et al., 2013; Albert, 2016). There are three clustering variables (j) in this analysis since initial velocity, horizontal ball movement, and vertical ball movement variables are used. This analysis uses a tuning parameter (λ) value of 0.5 since Pane et al. (2013) determined it was the optimal choice for reducing pitch misclassifications.

The first penalty factor used in the BIC adjustment is calculated by summing each intra-cluster correlation matrix. The combined logs of these sums are added together to create the first penalty (Equation 2). High values suggested there may be too many clusters given the large amount of cluster overlap (Pane et al., 2013). The second penalty factor accounts for the number of clusters, with preference given to models with smaller numbers of clusters (Equation 2).

The BIC value minus the two penalty values results in the BIC_{adj} value. The cluster configuration with the lowest BIC_{adj} is used for each pitcher. Clusters are deleted if they contain fewer than 10 pitches in a single year or fewer than one average pitch per game appearance.

Identifying outlier pitches in a model-based clustering method is especially important. Evans et al. (2015) stated, "Because little is known about the number or variance structure of the groups in the data *a priori*, model-based clustering methods are heavily data driven and results can be influenced by outliers" (p. 64). To identify outliers, the "FAST-MCD" method developed by Rousseeuw and Van Driessen (1999) is used to identify each cluster's mean using the minimum covariance determinant. The cov.mcd function in the R MASS package is used to create 97.5% Gaussian confidence ellipsoids for each cluster using the Mahalanobis distances (explained in more detail below) of all the points and their resulting covariance matrixes (Rousseeuw & Van Driessen, 1999; Hardin & Rocke, 2004; Hardin & Rocke, 2005). Pitches more than two standard deviations from each cluster's mean are removed as outliers. The percentage of pitches removed as outliers (Outlier %) is between 3–6% for most pitchers in the data set (Figure 5).

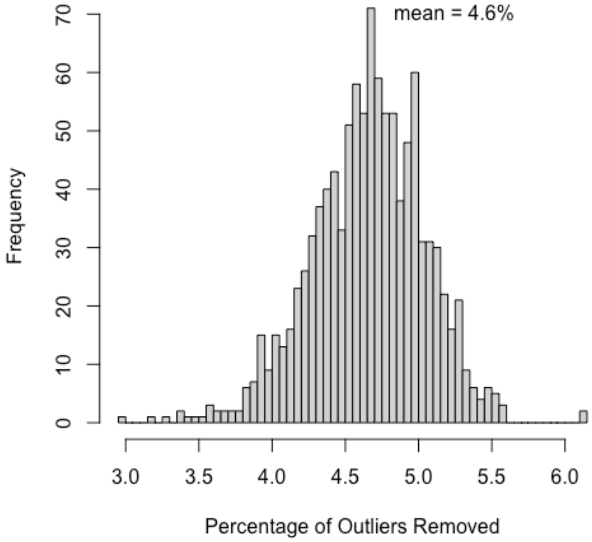

Figure 5. Percentage of pitches removed as outliers from observations in the data set.

Two pitchers in the training set and one pitcher in the test set have only two pitch clusters. These pitchers are removed from analysis due to their small frequency.

Table *1* lists the distribution of pitch types in the training and test sets. The majority of pitchers (59.5%) have either four or five pitch types. While only 15.2% of pitchers have more than six pitch types, this proportion is likely higher than the true proportion found in practice. For comparison, the PITCHf/x classification system identifies only 6.1% of the observations with more than six pitch types.

Table 1

Number of Pitchers with 3–9 Pitch Clusters in the Training and Test Sets

	Number of pitch clusters						
Data	3	4	5	6	7	8	9
Training	105	287	235	114	79	42	32
Test	32	72	39	18	5	2	2
Total	137	359	274	132	84	44	34

The relatively large number of observations with more than six pitch types may be due to several factors. Pitchers may encounter velocity or movement differences before and after an in-season injury, or when pitching at altitude, thereby making the same pitches appear in two distinct clusters. The most likely cause, however, is the PITCHf/x measurement discrepancies across stadiums—especially in earlier seasons when differences were more pronounced. 72.8% of the observations were in 2012 or 2013 compared to only 27.2% of the observations in 2014, 2015, 2016, or 2017. Moreover, the largest proportion of pitchers with more than six pitch types were with the Detroit Tigers. Compared to the rest of the league, Detroit had the largest PITCHf/x measurement discrepancy in 2013 (Roegele, 2013). Future iterations of this analysis may seek to control for these PITCHf/x park effects.

Distance Between Pitch Clusters

This analysis examines the relationship between pitches to identify factors affecting pitcher strikeout rates. One way to quantify differences between pitch types is to measure the distance between pitch clusters. The greater the distance, the greater the difference between pitches. Pitches separated by the largest distances are expected to deceive more batters.

Pitches clustered using velocity, horizontal movement, and vertical movement can be graphically represented on a 3D plot with each variable on a separate axis. (See Figure 4 above). One way to measure distances between these clusters is to use a common "straight-line" Euclidean measurement. Euclidean measurements ignore correlations between the variables, however. Gravity causes slower pitches to encounter more vertical movement than the same pitches at higher velocities. Velocity and vertical movement are highly correlated (r = .68). It is unusual for objects—such as curveballs and "rising" fastballs—to fall more or less than expected from gravity. The "difference" between two pitches should therefore incorporate the natural correlation between velocity and vertical movement to account for unexpected pitch movements.

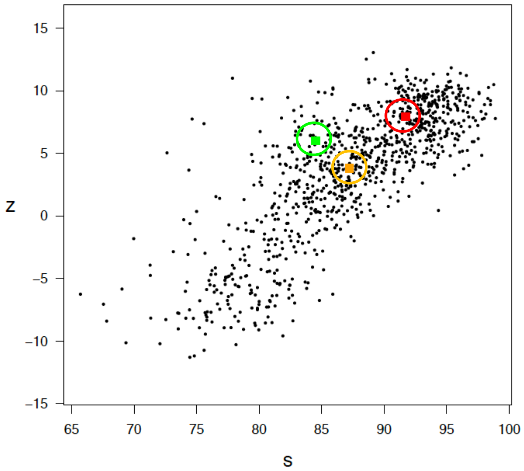

Figure 6. Mean pitch velocities and vertical movements for right-handed pitchers in 2016. Velocity (s) measured in miles per hour and vertical movement (z) measured in inches. Values for the colored points are described in Table 2. Reprinted with permission from Healey et al. (2017b).

Healey, Zhao, and Brooks (2017b) demonstrate why using Euclidean distances to measure differences between pitches can be misleading. Figure 6 is reprinted with permission from Professor Glenn Healey as it provides the clearest example of the drawbacks of using Euclidean distances to compare baseball pitches. Figure 6 shows the mean velocities and vertical movements of pitches thrown by right-handed pitchers in 2016. The green, orange, and red points represent three different pitch types. Table 2 provides a description of each pitch.

Table 2.

Details of the Three Colored Points in Figure 6.

Point Color	Pitcher	Pitch Type	Velocity (s)	Vertical Movement (z)
Green	Ian Kennedy	Changeup	84.51	6.01
Orange	Mat Latos	Cut Fastball	87.22	3.81
Red	Jhoulys Chacin	Four-Seam Fastball	91.71	7.94

Note. Reprinted with permission from Healey et al. (2017b). Velocity in miles per hour and vertical movement in inches.

Using the orange pitch as the reference point, the Euclidean distance to the red pitch (6.10) is greater than the distance to the green pitch (3.49). The larger distance to the red pitch suggests the orange and red pitches are very different from each other and that a combination of these pitches has a high likelihood of deceiving a batter. This assumption is likely incorrect.

A batter expects gravity to impart more downward movement on pitches at lower velocities. The difference between the red and orange pitches is therefore consistent with the correlation between velocity and vertical movement. While large in absolute/Euclidean distance from each other, a batter seeing both pitches would not be surprised because slower pitches tend to break more. It would be more surprising for the batter to see the green pitch. Although it has a lower velocity than the orange pitch, the green pitch's downward break is *less*. The batter would not expect the slower green pitch to seemingly defy gravity. In statistical terms, the green pitch is orthogonal to the orange pitch. Despite a small Euclidean distance between the orange and green pitches, the combination of these two pitches would be fairly deceptive to a batter.

Given these considerations, the Mahalanobis distance is a more reliable measure of the difference between pitch clusters. The Mahalanobis distance divides the standardized version of each value by the covariance matrix. The result is a unitless measurement accounting for correlations in the data (Mahalanobis, 1936; De Maesschalck, Jouan-Rimbaud, Massart, 2000). The Mahalanobis distance (*D*) for each point (*x*) is defined as:

$$D = (x-m)^{T}\ C^{-1}\ (x-m) \tag{3}$$

where *m* is the vector of mean variable values and C^{-1} is the inverse covariance matrix of the variables (De Maesschalck, Jouan-Rimbaud, Massart, 2000).

The Mahalanobis distance uses the covariance matrix for velocity, vertical movement, and horizontal movement to account for correlations between these variables and more accurately reflect differences between pitches. This measurement is therefore used to calculate differences between pitch clusters (Healey et al., 2017b). The Mahalanobis distances between the orange and red points is 0.81 units and the distance between the orange and green points is 1.32 units. From the orange point, the Mahalanobis distance to the green point is larger than the distance to the red point. The larger Mahalanobis distance between the orange and green points is consistent with intuition regarding the difference between these pitches.

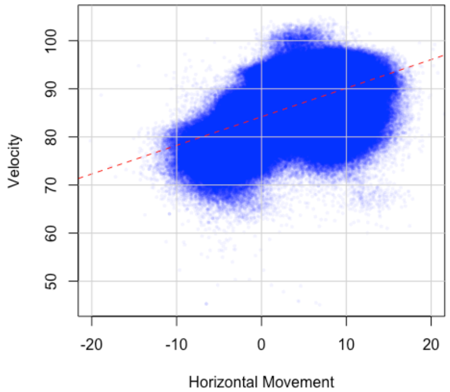

Figure 7. Pitch velocity and horizontal movement by left-handed pitchers from 2012–16.

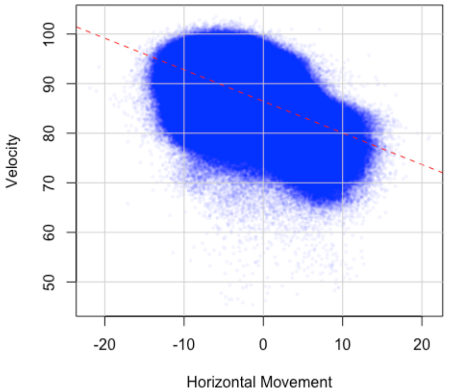

Figure 8. Pitch velocity and horizontal movement by right-handed pitchers from 2012–16.

When calculating Mahalanobis distances, separate correlation matrixes are used for left and right-handed pitchers. The opposite arm angles from these pitchers result in natural differences in horizontal movement. As shown in Figure 7 and Figure 8, correlations between velocity and horizontal movement are the opposite for left- and right-handed pitchers. Pitches by left-handed pitchers move left horizontally at lower speeds while pitches from right-handed pitchers move in the opposite direction.

Left and right-handed correlation matrixes are prepared by examining 982,024 and 2,566,791 pitches by left- and right-handed pitchers, respectively, from 2012–2016. As seen in Table 3, correlations between velocity and horizontal movement for left and right-handed pitchers have opposite signs. These correlation matrixes are used to calculate Mahalanobis distances between pairs of pitch clusters for left and right-handed pitchers. The pairwise.mahalanobis function in the HDMD package calculates mean pitch velocity, horizontal movement, and vertical movement for each cluster as well as the Mahalanobis distances between cluster means.

Table 3.

Velocity and Movement Correlation Matrixes for Left- and Right-Handed Pitchers

Left-Handed Pitchers				Right-Handed Pitchers			
	Velocity	Horizontal	Vertical		Velocity	Horizontal	Vertical
Velocity	1.000	0.523	0.687	Velocity	1.000	-0.574	0.685
Horizontal	0.523	1.000	0.523	Horizontal	-0.574	1.000	-0.519
Vertical	0.687	0.523	1.000	Vertical	0.685	-0.519	1.000

Mahalanobis distances are also used to measure the mean distance of all pitches from the center of their respective clusters (Precision). This Precision variable measures how consistently a pitcher can throw the same pitch type, with the hypothesis that less precision results in fewer strikeouts.

Independent Variables

Based on by findings from Gray (2002), Hale (2013), Arthur (2014), and Healey, Zhao, and Brooks (2017c), variables are created measuring minimum and maximum pitch type velocities, horizontal movements, vertical movements, and pitch release points. The range of these values among a pitcher's pitch types provides a measure of pitch variance—with larger variances indicating a broader range of velocity and movement. Another variance measurement calculates the size of the interquartile range (IQR) between the 25th and 75th percentiles for each of these variables. For each pitcher in this analysis, the size of the interquartile range is used to measure the variability of the above pitch metrics.

Variables measure maximum and average Mahalanobis distances between all pitch clusters. There are also two variables providing weighted measurements of the Mahalanobis distances among all pitch types. These weighted Mahalanobis variables attempt to aggregate the differences between pitch types into a single weighted measurement. The first variable aggregates the Mahalanobis distance of each pitch cluster from the center of the cluster of pitches thrown most frequently, weighted by pitch frequency (Weighted Mahalanobis Distance 1 or WMD-1):

$$\text{WMD-1} = \sum_{2}^{k} [D_{\text{k}}\ (n_{\text{k}}/n_{\text{total}})] \tag{4}$$

Where D = Mahalanobis distance from the center of the cluster of pitches thrown most frequently, k = number of clusters, and n = number of pitches.

A second variable sums the Mahalanobis distances between all pitch cluster pairs, weighted by pitch frequency (Weighted Mahalanobis Distance Between Pairs or WMD-BP):

$$\text{WMD-BP} = \sum_{i=1}^{k-1} \sum_{j=2}^{k} \sum_{i=1}^{n\text{-}1} \sum_{j=2}^{n} D_{\mu\text{i-}\mu\text{j}} \cdot [(n_i/n_{total}) + (n_j/n_{total})] \tag{5}$$

Where $D_{\mu i\text{-}\mu j}$ = Mahalanobis distance between the center of two clusters, k = number of clusters, i and j are cluster numbers, and n = number of pitches. For comparison, analogous weighted measurements are made using the Euclidean distance between pitch clusters and labeled WED-1 and WED-BP, respectively.

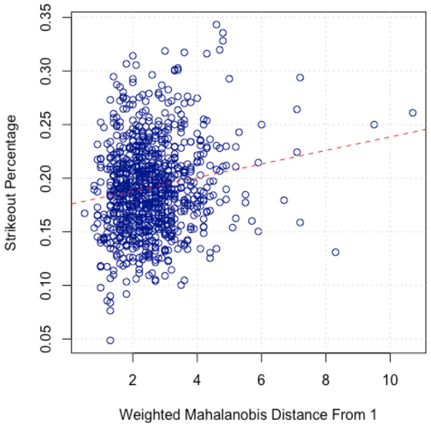

Figure 9. WMD-1 against strikeout percentage from 2012–16.

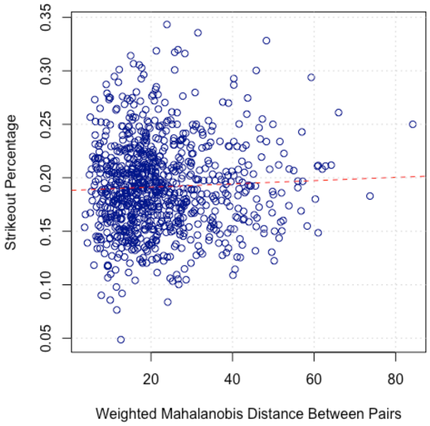

Figure 10. WMP-BP against strikeout percentage from 2012–16.

Figure 9 and Figure 10 show the relationship between strikeout percentage and each weighted aggregate Mahalanobis distance measurement. Strikeout percentage has a stronger correlation with WMD-1 (r = .16) than with WMD-BP (r = .04). The correlation between strikeout percentage and WMD-1 suggests that the most frequently thrown pitch serves to set-up the remaining pitches.

In addition, an entropy variable is created to measure uncertainty caused by the number of pitch types and the frequency with which each is thrown. The entropy equation introduced by Shannon (1948) is used to calculate each pitcher's entropy value:

$$\text{Entropy: } \sum_{i=1}^{k} \frac{n_i}{n_{total}} \left(log_2 \frac{n_i}{n_{total}} \right) \qquad (6)$$

Where n = number of pitches, k = number of clusters, and i is the cluster number.

Finally, a variable for pitcher strike percentage/rate is included. Strikeout percentage is highly correlated with strike rate (r = 0.30). Strike rate differences among pitchers can overshadow the impact of pitch movement and velocity on strikeout percentages. As this analysis seeks to isolate the impact of differences in velocity and movement on strikeout percentages, accounting for differences in strike rates addresses a confounding variable.

Table 4 provides summary statistics for each of the independent variables used in this analysis. The minimum, maximum, and range variables use the mean values of a pitcher's pitch type clusters rather than the absolute maximum and minimum values of all pitches thrown in a season. Conversely, the IQR variables examine all pitches thrown by a pitcher each season, regardless of the pitch type/cluster.

Table 4.

Statistics for the 894 Training Set Observations Used to Create Models

Variable	Mean	SD	Min	Max
Distance Between Pitch Types				
Average Mahalanobis distance	5.83	1.89	1.10	18.04
Maximum Mahalanobis distance (i.e., range)	11.79	5.46	1.08	46.43
WMD-1	2.57	1.10	0.50	10.80
WMD-BP	22.32	12.27	1.10	83.40
WED-1	19.69	10.88	3.20	55.30
WED-BP	2.60	0.89	0.90	5.80
Pitch Velocity				
Minimum	77.62	4.04	58.00	88.45
Maximum	91.51	2.43	81.86	98.15
Range	13.89	3.34	5.71	34.07
Velocity IQR (all pitches)	7.83	2.25	2.10	16.90
Pitch Horizontal Movement				
Minimum	1.66	1.38	0.00	8.15
Maximum	8.28	1.68	2.90	14.34
Range	6.62	2.04	0.98	13.26
Horizontal movement IQR (all pitches)	6.78	2.45	2.35	19.55
Pitch Vertical Movement				
Minimum	2.28	1.58	0.00	8.79
Maximum	9.22	1.82	3.06	17.73
Range	6.94	2.15	0.66	15.08
Vertical movement IQR (all pitches)	6.50	2.71	2.30	18.50
Pitch Pitcher Release Point				
Horizontal release point range	0.27	0.15	0.01	1.53
Vertical release point range	0.31	0.16	0.02	1.28
Horizontal release point IQR (all pitches)	0.34	0.13	0.17	1.34
Vertical release point IQR (all pitches)	0.24	0.07	0.14	1.11

Table 5 (continued).

Variable	Mean	SD	Min	Max
Other				
Number of pitch types	5.03	1.51	3.00	9.00
Entropy	1.96	0.42	0.92	3.09
Precision IQR	2.86	0.10	2.51	3.18
Strike percentage	0.64	0.02	0.55	0.73
Factor Variables				
League: American League = 421; National League = 427; Both = 46				
Horizontal release point IQR – High value flags = 21				

Note: Original values prior to log transformation.

Transformation and Flag Variables. Logarithmic transformations of independent variables are used to improve normality and increase prediction accuracy. Cross-validation within the training set reveals the lowest in-sample mean absolute prediction errors using independent variables with logarithmic transformations.

Indicator/flag variables corresponding to each independent variable are created to identify any values more than three standard deviations from each variable’s mean. Only flag variables with at least 2% of the observations outside of three standard deviations are included in the models to ensure a representative number of positive occurrences are present when modeling these factor variables. Consequently, only the horizontal release point IQR high flag is included when modeling.

Models Examined

Thirteen model types are evaluated using strikeout percentage as the dependent variable: multiple linear regression without (MLR) and with (MLR+) interaction terms, lasso and ridge shrinkage regression methods, principal components (PCR) and partial least squares (PLS) regressions, polynomial regression, regression splines, generalized additive models (GAM), random forests, boosted tree-based models, neural networks, and support vector machines (SVM). As the dependent variable is a proportion, predictions below zero or above one are capped to ensure predictions remain within an acceptable range. None of the predictions in the test set fell outside this range.

Stepwise AIC selection is used in the MLR, MLR+, polynomial regression, regression spline, and GAM models to identify independent variables. The MLR+ model started with interactions between six variables highly correlated with strikeout percentage: vertical movement IQR, vertical movement range, maximum vertical movement, strike rate, and velocity IQR. Variables with high multicollinearity are then removed from each model to ensure all variables in the final model have variable inflation factor (VIF) values less than 10. (See Black, 2014). In particular, none of the interaction terms remained in the MLR+ model after removing terms for multicollinearity.

The polynomial, spline, and GAM models use k-fold cross-validation prediction errors to identify the best iteration of each variable for modeling. Each variable in the polynomial model is tested against strikeout percentage in a simple linear regression model using one to five polynomial orders. The order of each variable with the lowest in-sample prediction error is used in the polynomial model. The spline model uses the same technique to identify the optimal number of degrees of freedom for either natural or smoothing spline variables. The GAM model uses either the polynomial, natural spline, smoothing spline, or local regression (LOESS) version of each variable with the lowest in-sample error.

K-fold cross-validation is used in several models to identify certain hyperparameters: the best lambda tuning values for the ridge and lasso regression models; the number of trees in the random forest model; the variable interaction depths and number of trees in the boosted model; the number of components used in the PCR and PLS models; the number of hidden layers to use in the neural net model; and the optimal cost and gamma values for the SVM model.

Results

Pitch Types

Figure 11 separates strikeout percentage by the number of pitch types. There is no apparent correlation between the number of pitch types and a pitcher's strikeout percentage (r = -.01). Gray (2002) and Arthur (2014) suggested pitchers with more than two pitch types have higher strikeout rates. While pitchers with three pitch types may have more success striking out batters compared to two pitch type pitchers, these benefits may dissipate as the number of pitch types increases. For many pitchers, it is possible the fourth, fifth, or sixth pitches in a pitcher's arsenal are not as effective as the first two or three.

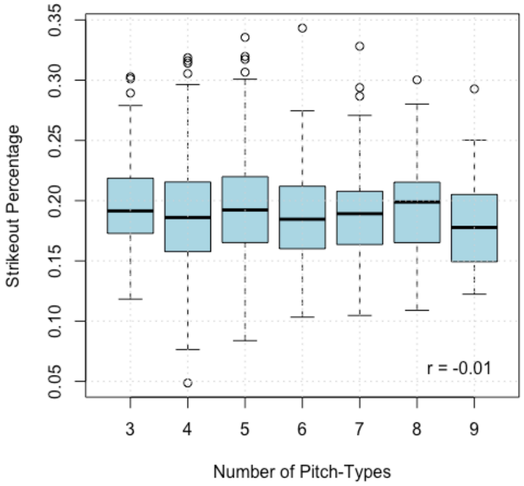

Figure 11. Strikeout percentage by number of pitch types from 2012−16.

Model Results

Statisticians have recognized that fastball velocity and strike rates affect pitchers' strikeout percentages. Combining these two variables serves as a control to measure the success of the present analysis over prior established techniques. A linear regression model is created from the training data with strike percentage as the dependent variable and both maximum pitch velocity and strike rate as independent variables. This model has an adjusted r-squared value of 0.230 and a MAE of 0.0331 against the test set.

Table 6

Test Set Errors for All Models

Model	MAE
Random Forest	0.0294
Multiple Linear Regression	0.0298
Principal Components Regression	0.0299
Boosted	0.0300
Ridge Regression	0.0300
Lasso Regression	0.0301
Regression Spline	0.0301
Multiple Linear Regression with Interaction Terms	0.0304
Partial Least Squares Regression	0.0305
Polynomial Regression	0.0308
Neural Network	0.0309
Generalized Additive Models	0.0310
Support Vector Machine	0.0324
Control Model: Strike Rate + Maximum Velocity	0.0331

Against the 2017 test set, the models in this analysis have MAE values between 0.0294 and 0.0324 (Table 6). The random forest model is the best performing model with a MAE of 0.0294, or 2.94%. This error is 0.0037 points (0.37%) less than the control model MAE.

Differences in pitch velocity and vertical movement have the greatest impact on strikeout percentage (Figure 12). The random forest model identifies these variables (in addition to strike rate) as the most influential in predicting strikeout percentage. These findings are consistent with those by Cameron (2009) and Arthur (2014) indicating that a pitcher's maximum pitch velocity is a strong determinant of his strikeout rate. Notably, one of the weighted Mahalanobis distance variables (WMD-1) is included in the 10 most important predictor variables. While vertical and horizontal pitch movements are important predictors of a pitcher's strikeout percentage, the distance between pitch types also appears to be a strong determinant of a pitcher's strikeout percentage.

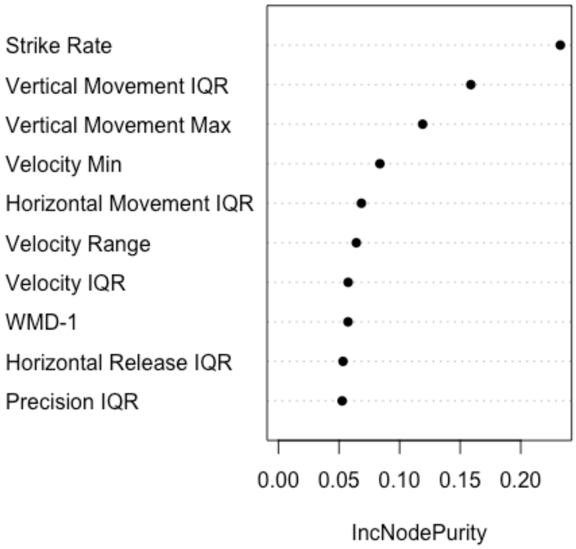

Figure 12. Random forest model variable importance plot (top ten variables).

Figure 13 shows predicted strikeout percentages against actual rates in the 2017 test set. Although the predicted distribution is narrower than the actual distribution, a large number of predictions are close to their actual values. Figure 14 shows the percentage of predictions captured within various error ranges. Of the 170 predictions against the test set, 147 (86.5%) are within five points of the actual strikeout percentages. Over half of the predictions are within 2.5

percentage points. The model's adjusted r-squared values in-sample and against the 2017 test set are 0.95 and 0.42, respectively.

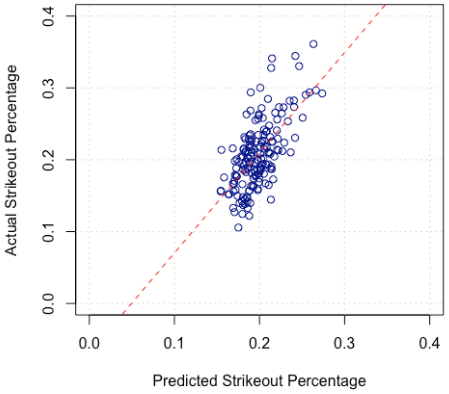

Figure 13. Random forest model 2017 predicted strikeout percentages against 2017 actual values.

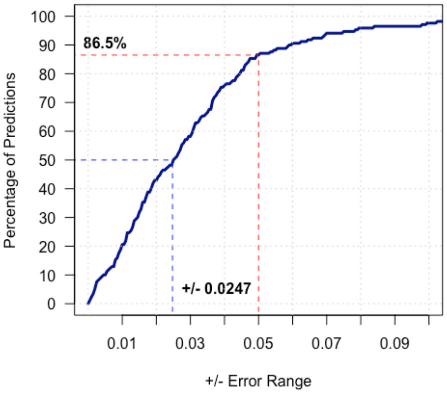

Figure 14. Percentage of random forest model predictions within error ranges.

Table 7

Top 10 Random Forest Model Predictions

Player	Pitch Types	2017 Strikeout Rates (%) Prediction	Actual	Absolute Error
Chad Kuhl	5	21.05	21.07	0.02
Cole Hamels	4	17.02	17.07	0.05
Jhoulys Chacin	5	20.20	20.13	0.07
Chris Stratton	5	19.84	19.92	0.08
Trevor Williams	8	18.40	18.28	0.12
Chad Bell	4	19.45	19.59	0.14
Zack Wheeler	5	21.16	20.98	0.18
Sam Gaviglio	3	15.42	15.61	0.18
Daniel Norris	4	18.64	18.82	0.18
Mike Leake	4	16.87	16.67	0.21

Table 7 lists the 10 predictions with the lowest errors from the 2017 test set. Each of these predictions has an absolute prediction error of 0.21 percentage points or less and include pitchers with 3–8 different pitch types.

Effect of MLB Strikeout Rate Increases on Prediction Errors

Prediction error rates are affected by the steady annual increase in MLB strikeout percentages (Figure 15). The MLB strikeout percentage increased from 16.6% in 2005 to 21.8% in 2017—a mean increase of 0.43 percentage points each year. As seen in Figure 15, the annual increase has been relatively constant.

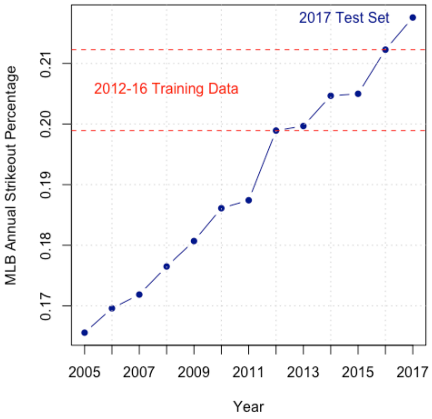

Figure 15. MLB annual strikeout percentages from 2005–2017.

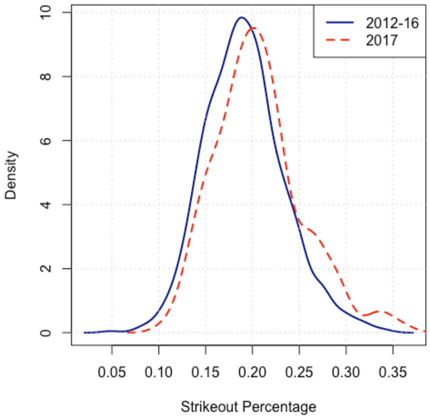

Figure 16. Distribution of strikeout percentages in the training (2012–16) and test (2017) sets.

Models in this analysis are trained on 2012–2016 data and tested on 2017 data. Figure 16 shows the difference in strikeout percentages between these data sets. The mean 2017 strikeout percentage (21.8%) was 2.6 percentage points higher than the mean combined 2012–2016 strikeout percentage (19.2%). While only 8.9% of pitchers in 2012–16 had strikeout percentages greater than 25%, almost twice that proportion (17.6%) exceeded 25% in 2017.

League-wide increases in strikeout percentage will adversely affect prediction accuracy if these increases are unrelated to the factors examined in this analysis. Table 8 shows the variables most strongly correlated with strikeout percentage and their compound annual growth rates (CAGR) from 2012 to 2017. While MLB strikeout percentages have increased 1.90% each year during the analysis period, none of the most influential variables in this analysis experienced similar growth. This suggests increases in MLB annual strikeout rates are unrelated to the variables examined in this analysis.[4] Including an adjustment for the estimated annual increase in the league strikeout percentage decreases the random forest model's MAE to 0.0288.

Table 8

Top Strikeout Percentage Predictor Variables

Variable	r	CAGR
Maximum Velocity	.377	0.05%
Strike Percentage	.300	-0.11%
Vertical Movement IQR	.255	-0.35%
Maximum Vertical Movement	.202	0.03%
Vertical Movement Range	.177	-0.55%
Minimum Velocity	.177	0.08%
Velocity IQR	.160	0.60%
WMD-1	.146	-0.92%

Note. CAGR of mean values for qualifying pitchers from 2012–17. Strikeout percentage CAGR was 1.90%.

[4] Commentators have provided a range of explanations for the increasing strikeout trend. In 2013, authors pointed to increases in both the number of pitchers used in games and batters' willingness to swing freely with two strikes (Carter, Quealy, & Ward, 2013). More recently, strikeout increases have been attributed to batters' efforts to increase their swing launch angles (Olney, 2017; Lindbergh, 2018). Regardless, such developments are unrelated to differences in velocity and movement among pitch types and are therefore not captured by this analysis.

Conclusions

The factors having the greatest impact on a pitcher's strikeout rate are his strike rate and vertical pitch movement. In addition, changes in velocity among pitches increase a pitcher's strikeout percentage. Horizontal movement and the Mahalanobis distance between pitches also impact the strikeout percentage.

Going forward, several additional analyses could build upon findings in this thesis. Research could determine whether factors affecting strikeout percentage differ based on the handedness of the batter and pitcher. Same-handed matchups typically favor pitchers. It would be interesting to understand whether differences in horizontal movement among pitches have a larger effect on strikeout percentage in these situations. Similarly, it would be helpful to understand whether there are changes in pitch entropy based on the handedness of the batter as pitchers may not throw certain pitch types based on the handedness of the batter.

Future work could also examine how differences in velocity and movement impact strikeout percentage each time through the batting order. Similarly, batter contact rates would also be an interesting predictor to add to future analyses. This could confirm whether differences in pitch velocity and movement have diminishing effects against batters with high contact rates. Moreover, while the present analysis focuses on starting and long-relief pitchers, future work could seek to understand whether the factors impacting strikeout percentage for situational and short-relief pitchers differ.

References

Adair, R. K. (1995). The physics of baseball. *Physics Today*, 48(5), 26-31. https://doi.org/10.1063/1.881460.

Albert, J. (2006). Pitching statistics, talent and luck, and the best strikeout seasons of all-time. *Journal of Quantitative Analysis in Sports*, 2(1), Article 2. https://doi.org/10.2202/1559-0410.1014.

Albert, J. (2016). Improved component predictions of batting and pitching measures. *Journal of Quantitative Analysis in Sports*, 12(2), 73-85. https://doi.org/10.1515/jqas-2015-0063.

Arthur, R. (2014, Feb. 6). Baseball proGUESTus: Entropy and the eephus. *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/22758/baseball-proguestus-entropy-and-the-eephus/.

Arthur, R. (2017, Apr. 28). Baseball's new pitch-tracking system is just a bit outside: As MLB switches from PITCHf/x to Statcast, the new tool is going through growing pains. *FiveThirtyEight*. Retrieved from https://fivethirtyeight.com/features/baseballs-new-pitch-tracking-system-is-just-a-bit-outside/.

Baumer, B.S., Jensen, S.T., & Matthews, G.J. (2015). openWAR: An open source system for evaluating overall player performance in major league baseball. *Journal of Quantitative Analysis in Sports*, 11(2), 69-84. https://doi.org/10.1515/jqas-2014-0098.

Branch, J. (2015, Oct. 4). Baseball talk, and all that stuff. *The New York Times* (New York print ed.), A1.

Bonney, P. (2015, Mar. 6). Defining the Pitch Sequencing Question. *The Hardball Times*. Retrieved from https://www.fangraphs.com/tht/defining-the-pitch-sequencing-question/.

Boyle, W., O'Rourke, S., Long, J. & Pavlidis, H. (2018, Jan. 29). Robo strike zone: It's not as simple as you think. *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/37347/robo-strike-zone-not-simple-think/.

Cameron, D. (2009, Feb. 17). Velocity and K/9. *Fangraphs*. Retrieved from https://www.fangraphs.com/blogs/velocity-and-k9/.

Cameron, D. (2017, Apr. 4). About all these velocity spikes. *Fangraphs*. Retrieved from https://www.fangraphs.com/blogs/about-all-these-velocity-spikes/.

Carleton, R. A. (2015, Feb. 3). Baseball therapy: The power of changing speeds. *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/25494/baseball-therapy-the-power-of-changing-speeds/

Carter, S., Quealy, K., & Ward, J. (2013, Mar. 23). Strikeouts on the rise. *The New York Times*. Retrieved from https://archive.nytimes.com/www.nytimes.com/interactive/2013/03/29/sports/baseball/Strikeouts-Are-Still-Soaring.html.

Chai, T. & Draxler, R.R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7, 1247-1250. http://doi.org/10.5194/gmd-7-1247-2014.

Davis, E. (2016, Aug. 11). Greg Maddux was a power pitcher despite the low velocity. *SB Nation Beyond the Boxscore*. Retrieved from https://www.beyondtheboxscore.com/2016/8/11/12423936/greg-maddux-velocity-finesse-power-pitcher-no-hope-for-batters.

De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L. (2000). Tutorial: The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50, 1–18.

Evans, K., Love, T., & Thurston, S. W. (2015). Outlier identification in model-based cluster analysis. *Journal of Classification*, 32, 63-84. http://doi.org/10.1007/s00357-015-9171-5.

Fast, M. (2007, Aug. 2). Glossary of the Gameday pitch fields. *Fast Balls*. Retrieved from https://fastballs.wordpress.com/2007/08/02/glossary-of-the-gameday-pitch-fields/.

Fast, M. (2010a). What the heck is PITCHf/x? The Hardball Times Baseball Annual 2010. Chicago, IL:ACTA Publications.

Fast, M. (2010b, June 17). The Internet cried a little when you wrote that on it. *The Hardball Times*. Retrieved from https://www.fangraphs.com/tht/the-internet-cried-a-little-when-you-wrote-that-on-it/.

Fast, M. (2011, Mar. 2). Spinning yarn: How accurate is PitchTrax? *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/13109/spinning-yarn-how-accurate-is-pitchtrax/.

Florio, J., & Shapiro, O. (2016, Oct. 4). How will rising temperatures change baseball? *The Atlantic*. Retrieved from https://www.theatlantic.com/science/archive/2016/10/how-will-rising-temperatures-change-to-baseball/502778/.

Fraley C., & Raftery A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97(458), 611-631.

Fraley, C., Raftery, A.E., Murphy, T.B., & Scrucca, L. (2012). mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation.

Technical Report No. 597, Department of Statistics, University of Washington. Retrieved from https://www.stat.washington.edu/sites/default/files/files/reports/2012/tr597.pdf.

Garik. (2011, Feb. 10). Being cautious with using Pitchf/x data to evaluate stuff: The case of Kyle Drabek. *SB Nation Beyond the Boxscore*. Retrieved from https://www.beyondtheboxscore.com/2011/2/10/1982529/being-cautious-with-using-pitchf-x-data-to-evaluate-stuff-the-case-of.

Gray, R. (2002). Behavior of college baseball players in a virtual batting task. *Journal of Experimental Psychology: Human Perception and Performance*. 28(5), 1131-1148. http://doi.org/10.1037//0096-1523.28.5.1131.

Hale, J. (2013, Oct. 30). Baseball ProGUESTus: Is speed enough?: A PITCHf/x look at the effect of fastball velocity and movement. Baseball Prospectus. Retrieved from https://www.baseballprospectus.com/news/article/22139/baseball-proguestus-is-speed-enough-a-pitchfx-look-at-the-effect-of-fastball-velocity-and-movement/.

Hardin, J. & Rocke, D.M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis*, 44, 625-638.

Hardin, J. & Rocke, D.M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14(4), 928-946. http://doi.org/10.1198/106186005X77685.

Healey, G., Zhao, S. & Brooks, D. (2017a, July 10). Measuring pitcher similarity. *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/32199/prospectus-feature-measuring-pitcher-similarity/.

Healey, G., Zhao, S. & Brooks, D. (2017b). Measuring pitcher similarity: Technical details. *viXra.org*. Retrieved from http://vixra.org/pdf/1705.0098v1.pdf.

Healey, G. & Zhao, S. (2017c). Using PITCHf/x to model the dependence of strikeout rate on the predictability of pitch sequences. *Journal of Sports Analytics*, 3, 93-101. http://doi.org/10.3233/JSA-170103.

Jackman, S. (2015). Pitch arsenal scores. *The Hardball Times*. Retrieved from https://www.fangraphs.com/tht/pitch-arsenal-scores/.

Kagan, D. (2009). The anatomy of a pitch: Doing physics with PITCHf/x data. *The Physics Teacher*, Vol. 47, October 2009. http://doi.org/10.1119/1.3225497.

Kagan, D. (2018, Jan. 23). The physics of RoboUmp. *The Hardball Times*. Retrieved from https://www.fangraphs.com/tht/the-physics-of-roboump/.

Law, K. (2017). *Smart baseball: The story behind the old stats that are ruining the game, the new ones that are running it, and the right way to think about baseball*. New York, NY: HarperCollins.

Lindbergh, B. (2018, Mar. 28). Baseball's three true outcomes are continuing their takeover in 2018. *The Ringer*. Retrieved from https://www.theringer.com/mlb/2018/3/28/17171162/spring-training-home-run-strikeout-stats-three-true-outcomes-trend.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*. 2(1), 49-55.

Marchi, M. (2011, Feb. 25). Fine tuning PITCHf/x location data. *The Hardball Times*. Retrieved from https://www.fangraphs.com/tht/fine-tuning-pitchf-x-location-data/.

Nathan, A. (2012, Oct. 21). Determining pitch movement from PITCHf/x data. *The Physics of Baseball*. Retrieved from http://baseball.physics.illinois.edu/Movement.pdf.

Nathan, A. (2013, Mar. 26). BP unfiltered: Is “late break” real? *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/19994/bp-unfiltered-is-late-break-real/.

Nathan, A. & Brooks, D. (2017, Apr. 5). Prospectus Feature: Estimating Release Point Using Gameday’s New Start-Speed. *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/31529/prospectus-feature-estimating-release-point-using-gamedays-new-start-speed/.

Office of the Commissioner of Major League Baseball. (2018). *Official Baseball Rules*. (2018 ed.). New York, NY. Retrieved from http://mlb.mlb.com/documents/0/8/0/268272080/2018-Official-Baseball-Rules.pdf.

Olney, B. (2017, June 4). Olney: Is the obsession with launch angle helping or hurting hitters? *ESPN*. Retrieved from http://www.espn.com/blog/buster-olney/insider/post/-/id/16792/olney-are-hitters-obsessed-with-launch-angle-helping-or-hurting-their-game.

Pane, M. A., Ventura, S.L., Steorts, R.C., & Thomas, A.C. (2013). Trouble with the curve: Improving MLB pitch classification. *arXiv:1304.1756v1 [stat.AP]*. Retrieved from https://arxiv.org/pdf/1304.1756.pdf.

Pavlidis, H., Judge, J., & Long, J. (2017, Jan. 24). Prospectus feature: Introducing pitch tunnels. *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/31030/prospectus-feature-introducing-pitch-tunnels/.

Perpetua, A. (2016, Sept. 15). Spin rates, swinging strikes, and an xSwStrk stat. *Rotographs*. Retrieved from https://www.fangraphs.com/fantasy/spin-rates-swinging-strikes-and-an-xswstrk-stat/.

Piette, J., Braunstein, A., McShane, B. B., & Jensen, S. T. (2010). A point-mass mixture random effects model for pitching metrics. *Journal of Quantitative Analysis in Sports* (Online), 6(3), Article 8. http://doi.org/10.2202/1559-0410.1237.

Rescan, A. (2017, Oct. 12). Kyle Hendricks's greatness is about more than control and command: His velocity-less success depends on the movement, too. *Beyond the Boxscore*. Retrieved from https://www.beyondtheboxscore.com/2017/10/12/16464244/kyle-hendricks-cubs-game-five-nlds-velocity-control-command-movement.

Richmond, M. (2015, Apr. 7). Pitches and stuff: The curveball. *SOSH, Boston Sports*. Retrieved from http://sonsofsamhorn.com/baseball/baseball-101/technique/basic-curveball/.

Roegele, J. (2013, Sept. 13). Basic 2013 PITCHf/x velocity park effects: Calculating basic PITCHf/x velocity park effects and discussing the sources of error that are inherent in the numbers. *SB Nation Beyond the Boxscore*. Retrieved from https://www.beyondtheboxscore.com/2013/9/13/4720852/basic-2013-pitchfx-velocity-park-effects-error-sabermetrics.

Roegele, J. (2014, Nov. 24). The effects of pitch sequencing. *The Hardball Times*. Retrieved from https://www.fangraphs.com/tht/the-effects-of-pitch-sequencing/.

Rousseeuw, P., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212-223.

Ryan, N., & House, T. (1991). *Nolan Ryan's pitcher's bible: The ultimate guide to power, precision, and long-term performance*. New York, NY: Simon & Schuster.

Sarris, E. (2014, Dec. 16). Toward a pitching arsenal score statistic. *Rotographs*. Retrieved from https://www.fangraphs.com/fantasy/toward-a-pitch-arsenal-score-ranking-statistic/.

Sarris, E. (2018, Jan. 23). What Jack Flaherty has in common with Clayton Kershaw. *Fangraphs*. Retrieved from https://www.fangraphs.com/blogs/what-jack-flaherty-has-in-common-with-clayton-kershaw/.

Schmidt, M. & Ellis, R. (1994). *The Mike Schmidt Study: Hitting Theory, Skills and Technique.* Atlanta: McGriff and Bell.

Schwartz, D. (2014, Dec. 19). Pitch arsenal score part deux. *Rotographs*. Retrieved from https://www.fangraphs.com/fantasy/pitch-arsenal-score-part-deux/.

Shannon, C. (1948a). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3), 379-423, 27(4), 623-656, July, October, 1948.

Shannon, C. (1948b). A Mathematical Theory of Communication (continued). *The Bell System Technical Journal*, 27(4), 623-656.

Sidle, G. & Tran, H. (2018). Using multi-class classification methods to predict baseball pitch types. *Journal of Sports Analytics*, 4(1), 85-93.

Sievert, C. (2014). Taming PITCHf/x data with XML2R and PitchRx. *The R Journal*, 6(1),5-19.

Sports Reference LLC. (2018). *Baseball-Reference.com - Major League Statistics and Information*. Retrieved from https://www.baseball-reference.com/.

Trueblood, M. (2018, Feb. 1). Rubbing mud: The Cubs have already mined these tunnels. *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/37461/rubbing-mud-cubs-already-mined-tunnels/.

Willmott, C. & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res., 30, 79–82.

Willmott, C. J., Matsuura, K., & Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. Atmos. Environ., 43, 749–752, 2009.

Woolner, K. & Perry, D. (2006). Why are pitchers so unpredictable? [In] *Baseball Between the Numbers, Why everything you know about the game is wrong* (pp. 48-57). Keri, J. (ed.). Basic Books:New York, NY.